DOCUMENT RESUME

ED 411 265 TM 027 215

AUTHOR Thomasson, Gary L.

TITLE The Goal of Equity within and between Computerized Adaptive Tests and Paper and Pencil Forms.

PUB DATE 1997-03-00

NOTE 25p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, March 25-27, 1997).

PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Ability; *Adaptive Testing; Comparative Analysis; *Computer Assisted Testing; Scores; Simulation; *Test Bias; Test Content; Test Format; Test Items

IDENTIFIERS *Paper and Pencil Tests

ABSTRACT

Score comparability is important to those who take tests and those who use them. One important concept related to test score comparability is that of "equity," which is defined as existing when examinees are indifferent as to which of two alternate forms of a test they would prefer to take. By their nature, computerized adaptive tests (CAT) add extra complexity for score comparability issues. At a minimum, comparability between a CAT and a paper-and-pencil (P&P) form means that the score each examinee gets on the CAT and P&P forms should be as interchangeable as possible in terms of the construct being measured and the measurement precision of that construct. The importance of the content issue, termed "content balancing," has been expressed by several researchers. The relative importance or appropriateness of content areas to different regions of the ability scale should guide development of both CAT and P&P forms. Among the approaches suggested for content balancing is a conditional balancing approach advocated by T. C. Davey and L. Thomassen (1995). The simulation study described in this paper extends Davey's approach by including realistic item exposure controls. An artificial 25-item P&P math test was used as a reference and target test for the CAT conditions. Results from this preliminary study seem to confirm the assertion that it is possible to employ item content controls appropriately during item administration of a CAT, following targets based on natural distribution of item content as a function of ability so that total measurement precision is not severely negatively impacted. (Contains 21 figures and 20 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 411 265

The Goal of Equity *within* and *between* Computerized Adaptive Tests and Paper and Pencil Forms

Gary L. Thomasson
Defense Manpower Data Center

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Gary Thomasson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Presented at the annual meeting of the National Council on Measurement in Education in Chicago, IL, March, 1997

The views and opinions contained in this report are those of the author and should not be construed as representing any official position or policy of any governmental agency.

Address correspondence to: Gary L. Thomasson
Defense Manpower Data Center
400 Gigling Road
Seaside, CA 93955-6771
Email: ThomasGL@osd.pentagon.mil
Email: thomasson@nps.navy.mil

TM027215

BEST COPY AVAILABLE

The Goal of Equity *within* and *between* Computerized Adaptive Tests and Paper and Pencil Forms

It is important to produce alternate forms of a test such that scores on one form are as comparable as possible to scores on an alternate form. Score comparability or score interchangeability is important both to examinees and to those who use test scores. Score comparability underlies the procedures of test equating and is an important component in test fairness issues. One important concept related to test score comparability is that of "equity[1]" proposed by Lord (1980) who defined equity as existing when examinees are indifferent as to which of two alternate forms of a test they would prefer to take.

Most psychometricians use the concept of equity as a desirable goal for alternate test form development and test equating, but one that cannot be perfectly attained in most realistic situations. At the examinee level, no two tests with different items can be truly equitable in the sense that each examinee would be completely indifferent. If the items differ in any way across two tests, we can almost assuredly find at least one examinee who would prefer one test over another. Even Lord (1980) mathematically describes equity as operating at the conditional population level, not the individual level; that is, the conditional score distributions should be equal across alternate test forms for any ability level. Others (e.g., Sympson, 1985; Thomasson, 1991a; 1991b; 1993a; 1993b) have extended Lord's concept to the moments of the conditional distributions and introduced terms such as first-order equity, second-order equity, higher-order equity, full-equity and local full-equity. However, definitions of equity based only on conditional distributions are meaningful only in a purely theoretical sense, that is, for theoretically pure tests for which scores are solely determined by a known finite-dimensional latent variable (commonly called theta). For real tests, we must always assume that known and unknown minor factors or nuisance variables or traits can effect scores on these real tests. Thus, fairness in testing would seem to argue for making tests, as much as possible, approach the equity criteria at the individual level as well as the conditional population level. Equity at the individual level is important to the examinee as well as to the users of these real tests who must make decisions based on test scores.

P&P forms:

For paper-and-pencil (P&P) test forms, the major test development tool that promotes comparable alternate forms is the test specification (e.g., Millman & Greene, 1989). This is the blueprint used to construct alternate test forms. The designing of forms using test specifications is critical even before one can consider test equating procedures (Kolen & Brennan, 1995).

---

[1] The term "equity" in testing has taken on many dimensions in recent years. The term as used in this paper is based on the usage by Lord (1980) and used in the context of equating alternate test forms. Meanings of "equity" beyond this context will not be directly addressed here.

Millman and Greene (1989) claim, "Foremost among the attributes of a test requiring specification is its content. Other important attributes include item types, psychometric characteristics, scoring criteria and procedures, and number of items to be developed." The level of content specification should be as detailed as necessary and practical to produce similar tests. Among other important test attributes, the test specifications generally denote how many items from each of several major content areas are to be administer in each fixed length P&P test form (along with other constraints). In this way alternate P&P forms can be said to measure the same content areas across forms. Assuming that alternate forms are produced that match each other exactly at a certain level of content specification, then if examinees consider only this level of content in evaluating the forms, they would be indifferent to such alternate forms.

CAT forms:

By their nature, computerized adaptive tests (CAT) add extra complexity for score comparability issues. The CAT item-selection algorithm must select items to administer on the basis of several competing goals. At least three major goals of a CAT item selection algorithm are (1) measurement precision and efficiency, (2) item security (or item exposure control to avoid overexposure of the most highly informative items), and (3) test specification constraints - including item content constraints. The advantage, as well as the disadvantage, of computerized adaptive testing is that examinees taking the same CAT form will be taking different subsets of items. One can view the CAT software as interacting with each examinee's responses to generate a personalized alternate test form for that examinee as he or she is taking the test. Thus, CAT administration software is functionally an on-line, personalized form-assembly program.

At a minimum, comparability between a CAT and a P&P form means that the score each examinee gets from the CAT form and the P&P form should be as interchangeable as possible in terms of the important issues of (1) the construct being measured (which is largely determined by the content classifications of the items administered) and (2) the measurement precision of that construct. And, comparability within a CAT form means that the score each examinee gets from (hypothetically) retesting on the same CAT test (and being administered possibly different items during each retest) should also be as interchangeable as possible in terms of the same two important issues of (1) the construct being measured (again, the content of the items administered) and (2) the measurement precision. The importance of the content issue (often termed "content balancing") in CAT has been expressed by several researchers (APA, 1986; Green, et al, 1984; Wainer, et al, 1990). Controlling the content of a CAT can be done a two levels. At the item pool level, we can control the numbers and quality of items from the important content areas to be placed in a CAT item pool; this is important since a CAT can only administer items that currently exist in its pool. However, at the level of item selection or item administration, there can also be controls over the content of administered items at the actual time of CAT administration.

Before considering content control in more depth, let's consider the importance of the conditional test construct in the control of item content.

Test construct as a conditional concept

With the tools of item response theory (IRT) becoming more and more available, Thomasson (1996) has argued that the test construct should be considered a concept that is conditional on ability level. This becomes especially important when the content of a test varies by difficulty level. For example, consider that the content areas of Pre-algebra, Algebra, and Geometry have been specified or chosen to represent the domain of a certain Math test. In such cases, the test construct changes in terms of the relative importance of the content areas as one moves along the ability scale. This has implications for test specification and for the construction of P&P as well as CAT test forms. The relative importance or appropriateness of content areas to different regions of the ability scale should guide the development of both P&P and CAT test forms.

Consider further this example where the test content varies by difficulty. Figure 1 shows the total test information function as well as content-specific test information functions for a hypothetical 25-item Math test composed of 7 Pre-algebra items, 9 Algebra items, and 9 Geometry items. Note that the relative contribution of each content areas to the total test information varies along the ability scale in a natural order. Pre-algebra is a relatively more important contributor to the test construct and to the test information of this Math test for the lower ability levels than for the higher ability levels. As one moves up the ability scale, Algebra becomes a more important contributor, and then Geometry becomes a more important contributor to the test construct and to the test information of this Math test for higher ability examinees. The implications for a comparable CAT test form should be that examinees of lower ability should be tested predominantly on content that is most appropriate for their ability range (i.e., Pre-algebra), while examinees of higher ability should be tested predominantly on content that is most appropriate for their ability range (i.e., Algebra and Geometry). Now let's consider how to best implement such a content-regulating idea.

Content-balancing in CAT: How?

Several researchers have proposed approaches and methods aimed at satisfying some compromise among two or more of the above mentioned, three major competing goals of a CAT (Davey & Thomas, 1996; Segall, & Moreno, 1986; Stocking & Swanson, 1993; Stocking & Lewis, 1995; Sympson & Hetter, 1985; Thomasson, 1995). This paper is interested in an approach similar to that taken by Tim Davey (Davey & Thomas, 1996) which was aimed at integrating the goals of measurement precision and of content balancing. Davey (Segall & Davey, 1995; Davey & Thomas, 1996) evaluated several methods of content-balancing a CAT version of the ACT Assessment Program Mathematics Test (AAP Math). Combining these studies, he evaluated the performance of the CAT test under the following conditions: (1) no content balancing (item selection based on item information only), (2) balancing fixed proportions of items based on fixed

target content percentages of items derived from the P&P form of test, (3) balancing fixed proportions of information based on these same fixed target content percentages of items on the P&P form of test, and (4) balancing the proportions of information based on target content percentages of total test information that are conditional on estimated ability (theta-hat). Davey's finding was that the last approach (4) produced content-specific information functions that best matched those of the target P&P form. As Davey indicated, further research is needed in several areas. In these studies Davey did not include any item exposure control nor did he look at variability of individual CAT tests presented conditioned on ability.

The actual feasibility of these content-balancing options varies. While it is almost always feasible to administer items under the first two options above, consider the third option when content domains vary with difficulty, as in Figure 1. It may not even be doable or feasible to administer items to very low ability examinees such that the low ability conditional information function contains, say, one-third of its total test information from the Geometry content area, since Geometry items "naturally" have more information at higher ability levels as compared with Pre-algebra items. Likewise, it may be nearly impossible to force very high ability conditional information functions to contain one-third of its test information from the Pre-algebra content-area, since Pre-algebra items "naturally" have more information at lower ability levels.

One purpose of the simulation study in this paper was to replicate and to extend Davey's approach by including realistic item exposure controls (Thomasson, 1995). Another motivation was to investigate the extent to which total test information is reduced by different methods of content balancing. Of special interest was the examination of the conditional variability of administered CAT forms in terms of the two issues of (a) the construct being measured (i.e., item content) and (b) the conditional measurement precision.

In the CAT program for the Armed Services Vocational Aptitude Battery (ASVAB) there have been three general approaches to content-related item selection implemented to date. These are (1) to ignore content during item selection (for most of the power tests), (2) to use an allocation vector that predetermines the content of each item during administration (for the General Science test), and, in the extreme case, (3) to partition a single P&P test into two separate CAT subtests (as in the case of P&P Auto-Shop test becoming separate Auto Information and Shop Information CATs). One reason for the reluctance to implement stronger content-related item selection on most power tests is that measurement precision would likely be lowered to undesirable levels while requiring rigid global content-related constraints during item administration. Based on the results from Davey cited above, it seems that some method of conditional content balancing could preserve adequate levels of measurement precision while at the same time controlling item-content during administration.

Methods

The main purpose of this study was to demonstrate and evaluate the item-selection conditions of (1) no content balancing (ignoring content during item selection), (2) to use an allocation vector that predetermines the content of each item during administration, and (3) a method of conditional content balancing based on target P&P content-specific information functions.

Unidimensionality

This preliminary study will assume that content areas are very highly correlated with the major dominant dimension of the test, and that the test can be adequately modeled by a single unidimensional latent trait. Even though the content areas are considered important to defining the test construct, the test will be assumed to behave statistically as an essentially unidimensional test. Thus, information functions for each content area can be meaningfully plotted on the same theta scale.

The P&P target test

An artificial 25-item P&P Math test was selected as a reference and target test form for the CAT conditions. This artificial Math test was composed of 7 Pre-algebra items, 9 Algebra items, and 9 Geometry items. The content-specific information functions and the total test information function for this P&P target test are shown in Figure 1. To a large extent, these target content-specific information functions are very similar in shape to the average content-specific information function of a large pool of these item types (See Figure 2), and thus may be said to represent a rather "natural" target for the content-specific information functions.

The CAT test

The CAT test used in this study follows the typical 15-item fixed length CAT ASVAB power test with a few exceptions. One exception was the use of an extra large item pool to eliminate basing the conclusions on a restricted item pool. The other exception was varying the method of item selection as the major conditions to be studied. The exposure control procedure, a modified Sympson-Hetter procedure with a 0.7 ceiling on all item exposure indices (Thomasson, 1995), was held constant across all conditions. The intermediate ability estimates are Owen's Bayes (1969) theta estimates.

The "no balance" item-selection condition

In this condition, the current ability estimate is used to select the next best item (highest conditional information at the current estimate and not previously used or considered for use) from an information table based on the full item pool, ignoring item-content. The exposure control index was used to decide whether to actually use this item or select the next best item from the information table.

The "allocation vector" item-selection condition

In this condition, an allocation vector was used to predetermine the item content of each item of the 15-item CAT. The allocation vector used was [P, A, G, P, A, G, P, A,

G, P, A, G, P, A, G]. That is, the first item was always selected from the Pre-algebra content, the second from the Algebra content, and the third from the Geometry content. This sequence was repeated until five item of each content area, 15 items total, was given. Like the "no balance" condition, the current ability estimate is used to select the next best item from an information table, but, in addition, the item-selection was conditional on item content as determined by the allocation vector. The same exposure control procedure is then consulted as to the actual use of this selected item.

The "minimum target conditional information" item-selection condition

In this condition, the item-content was determined by considering the content-specific target information conditional on the current estimate of ability rather than a predetermined allocation vector; a cumulative information function, conditional on item-content, is computed on the basis of each administered item. When the next item is to be selected, the intermediate ability estimate (Owen's Bayes) is computed and updated first. Then the cumulative information function, conditional on each item-content, is compared with the corresponding target conditional information at the point nearest the intermediate ability estimate, and a difference for each content is computed.

$$diff(\text{content}) = I_{\text{cumulative}}(\hat{\theta}, \text{content}) - I_{\text{target}}(\hat{\theta}, \text{content})$$

The content showing the largest negative difference with its conditional target is selected as the content of the next item to be administered. Then, as in the other conditions, the current ability estimate is used to select the next best item from an information table (subject to the same exposure controls), but the item-selection is conditional on item-content as determined by the largest negative *diff*(content). If none of the contents have a negative *diff*(content), then content is ignored and the most informative item, ignoring content, is selected (subject to exposure controls). In this way, the content-specific target information curves perform as minimum targets for the CAT content-specific information functions.

Note that this item selection algorithm differs from that used by Davey (Davey & Thomas, 1996) in his "balancing proportions of conditional information" design condition, however both algorithm adaptively select item content based on the current estimate of examinees ability and the content-specific P&P-based target information functions on theta.

Results

Total test information for each of the three conditions was plotted in Figure 3. In this example, using a rather large item pool, the impact of content balancing on total test information was relatively small and in the direction that one would predict. The greatest loss of total test information occurred with the allocation vector condition, which produced somewhat lower test information over the full range of abilities. The target

condition produced slightly lower total test information only where the no-balance condition produces very high information -- i.e., in the region (0.5 < theta < 1.8). It is hypothesized that the decrements in total test information would be relatively larger for smaller item pool sizes, but that the general pattern of information loss would be about the same.

More important for this study is the investigation of content-related information functions. Consider the plots of the mean CAT Geometry information functions by condition in Figure 4, and the corresponding plot of mean CAT Geometry information functions by condition minus the P&P Geometry target information function (Figure 5). While the Geometry information functions under both the allocation condition and the target condition were both above the minimum P&P Geometry information target for the almost all abilities, the no-balance condition produced Geometry information functions significantly lower than the P&P target in the theta range of (-0.4 through +1.2). In this range the content-related equity is severely impaired for the no-balance condition. Obviously, in these ranges, there are enough Algebra and Pre-algebra items with higher information than that of most of the "most informative" Geometry items, such that the Algebra and Pre-algebra items are selected ahead of the Geometry items in this region. Evidence for this can be seen in Figure 6, the graphs for Algebra information functions; in Figure 7, the graphs of Algebra information functions minus the P&P target; in Figure 8, the graphs for Pre-algebra information functions; and in Figure 9, the graphs of Pre-algebra information functions minus the P&P target.

General conclusions drawn from Figures 4 through 9 are that the mean content-specific information function from the target CAT condition always exceeded the P&P target information for each content area, but that the mean content-specific information function from the no-balance CAT condition sometimes fell quite short of the P&P target information. In general, the mean content-specific information function from the allocation-vector CAT condition exceeded the P&P target information for each content area, with the minor exception that for high abilities the Geometry information function fell slightly below that of the P&P target.

To examine the *within*-CAT equity or *within*-CAT variability in information functions, graphs of the conditional standard error of information functions were considered. Figures 10 through 12 show the mean information plus and minus one standard error for the target condition for the Geometry, Algebra, and Pre-algebra content areas respectively. Figures 13 through 15 graph the conditional standard error for each of the three conditions for the Geometry, Algebra, and Pre-algebra content areas respectively. In general it can be seen that the standard error of information is typically higher in regions where the mean information is higher. For the target condition, the P&P minimum target information is generally below the "mean minus one standard error" curve. This indicates that for the target condition, not only does the mean information exceed the P&P minimum target, but that even with variability in administered items, most examinees were administered CAT test forms that exceeded the minimum targets.

Other general conclusions about *within*-CAT equity that can be seen in Figures 13 through 15 are that the no-balance CAT condition was generally the most variable and that the allocation-vector CAT condition was generally the least variable. The target CAT condition might have been less variable than observed in these results if there were more constraints during the "no-balance" part of the target algorithm after the minimum target was satisfied.

The number of items from each content area was predetermined only by the allocation vector condition. For the other two CAT conditions, the number of items from each content area was not of direct concern (except the minimum number of one item from each content area for the target CAT condition). However, it may be of some interest to observe after the fact, how many items of each content area were administered by these two conditions. For the Geometry content area, Figure 16 displays the conditional average number of items given, plus and minus one standard error, for the target CAT condition, and Figure 17 displays the same curves for the no-balance condition. For the Algebra content area, Figure 18 displays the conditional average number of items given plus and minus one standard error for the target CAT condition, and Figure 19 displays the same curves for the no-balance condition. And, for the Pre-algebra content area, Figure 20 displays the conditional average number of items given, plus and minus one standard error, for the target CAT condition, and Figure 21 displays the same curves for the no-balance condition. The general trends in these graphs fit well with the expectation that low-ability examinees are administered more Pre-algebra items and higher-ability examinees are administered more Algebra and then Geometry items. In addition, the problematic drop in the Geometry content over the theta range of (-0.4 through +1.2) can be seen in the no-balance condition.

Discussion

The results of this preliminary study seem to confirm the assertion that it is possible to appropriately employ item content controls during item administration of a CAT, following targets based on natural distribution of item content as a function of ability in such a way that the total measurement precision is not severely negatively impacted. It is important not to destroy the sound validity and reliability of measurement instruments just to meet arbitrary globally-fixed "face validity" constraints or arbitrary global content constraints. This study demonstrates that if CAT item pools are sufficient, and that item selection constraints are appropriate, the sacrifice in measurement precision (and corresponding reduction in test reliability and a likely reduction in test validity due to an increase in measurement error) may be trivial. In addition, by using natural, appropriate constraints there may be gains in within-CAT and between-CAT and P&P equity.

More research is suggested to fine-tune the target algorithm for CAT and to apply it to more realistic situations. Next steps include using more realistically sized item pools and investigations with actual ASVAB test and targets.

Implication for test specifications and test blueprints

Although there will always exist multiple "solutions" to computerized testing that depend on a large number of psychometric, practical, and policy constraints (including such things as "market pressures" and various client requirements), some general desiderata are asserted here.

- Target content-specific information functions may evolve to become a standard part of the test specifications for CAT administration and for P&P form assembly
-
- Any setting of such content-specific information targets should be as "natural" as possible rather than arbitrary:
    - they should follow the natural shape of typical content-specific item pools, and
    - they should be feasible to obtain in following typical item development and CAT pool development efforts.
    -
- There may be continuing residual concern about too much information in a CAT at certain ability ranges with respect to corresponding P&P forms:
    - how to handle equity differences between administrations of CAT and P&P (due to the natural excess precision of CAT in some ability regions) without sacrificing the benefits of CAT will be an ongoing debate. Deliberately "crippling" a CAT to look like a P&P form seems to be an unacceptable option.

In conclusion, the method of conditional content balancing should provide researchers and practitioners with a CAT algorithm that more closely achieves the goals of CAT including equity *within*-CAT at the individual level.

References

American Psychological Association. (1986). Guidelines for computer-based tests and interpretations. Washington, DC: Author.

Davey, T. C., & Thomas, L. (1996, April). Constructing adaptive tests to parallel conventional programs. Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement, 21, 347-360.

Kolen, M. J., & Brennan, R. L. (1995). Test equating: Methods and practices. New York: Springer-Verlag.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Millman, J., & Greene, J. (1989). The specification and development of test of achievement and ability. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 335-366). New York: American Council on Education, Macmillan.

Owen, R. J. (1969). A Bayesian approach to tailored testing (RB-69-92). Princeton, NJ: Educational Testing Service.

Segall, D. O., & Moreno, K. E. (1986). Dimensionality for the ACAP item pools: Finding and recommendations. San Diego, CA: Paper presented at the meeting of the CAT-ASVAB technical committee.

Segall, D. O., & Davey, T. (1995, June). Some new methods for content balancing adaptive tests. Paper presented at the Annual Meeting of the Psychometric Society, Minneapolis, MN.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. Applied Psychological Measurement, 17, 277-292.

Stocking, M. L., & Lewis, C. (1995, August). A new method of controlling item exposure in computerized adaptive testing. Princeton, NJ: Educational Testing Service.

Sympson, J. B. (1985, August). Alternative objectives in test equating: Different goals imply different scales. Paper presented at the annual meeting American Psychological Association.

Sympson, J. B., & Hetter, R. D. (1985, October). Controlling item exposure rates in computerized adaptive testing. Proceedings of the 27th annual meeting of the Military Testing Association (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Thomasson, G. L. (1991, May). Equating tests in nonequivalent groups. Paper presented at the 1991 Office of Naval Research Contractors' Workshop on Model-Based Psychological Measurement, Princeton, NJ.

Thomasson, G. L. (1991, June). Equating tests in nonequivalent groups: An evaluation of frequency estimation. Paper presented at the Annual Meeting of the Psychometric Society, Brunswick, NJ.

Thomasson, G. L. (1993, June). Asymptotic equating methodology. Paper presented at the Annual Meeting of the Psychometric Society, Berkeley, CA.

Thomasson, G. L. (1993, October). Asymptotic equating methodology and other test equating evaluation procedures. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Thomasson, G. L. (1995, June). New item exposure control algorithms for computerized adaptive testing. Paper presented at the Annual Meeting of the Psychometric Society, Minneapolis, MN.

Thomasson, G. L. (1996, June). Suggested revised guidelines for "content balancing" in computerized adaptive testing. Paper presented at the Annual Meeting of the Psychometric Society, Banff, Alberta, Canada

Wainer, H., Dorans, N. J., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). Future challenges. In H. Wainer (Ed.), Computer adaptive testing: A primmer (pp. 233-286). Hillsdale, NJ: Lawrence Erlbaum Associates.

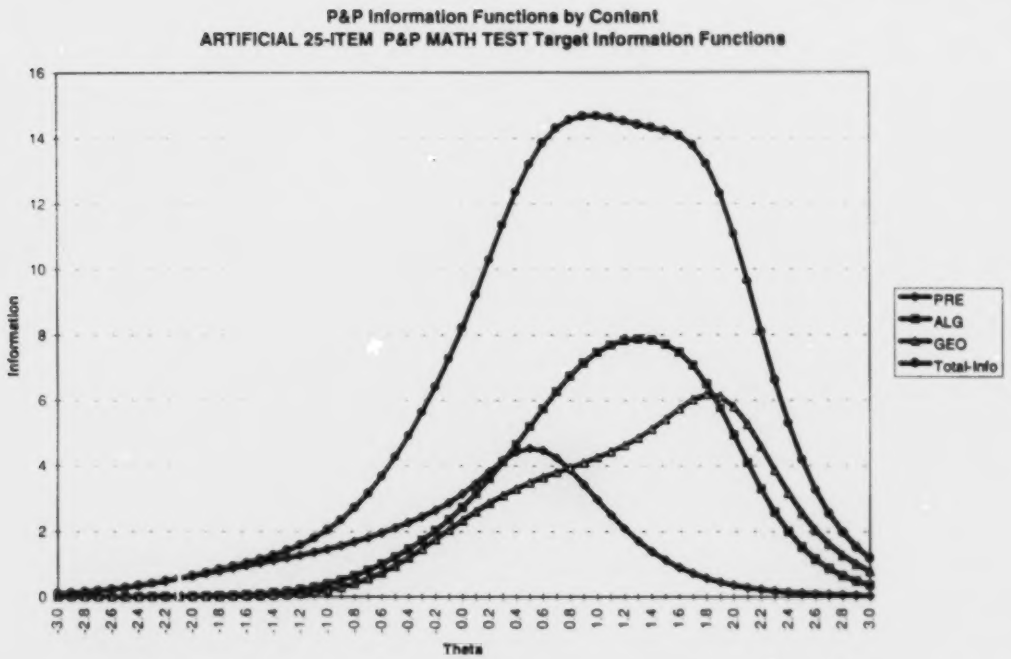

Figure 1. Information functions, total and by content, for artificial 25-item P&P Math test

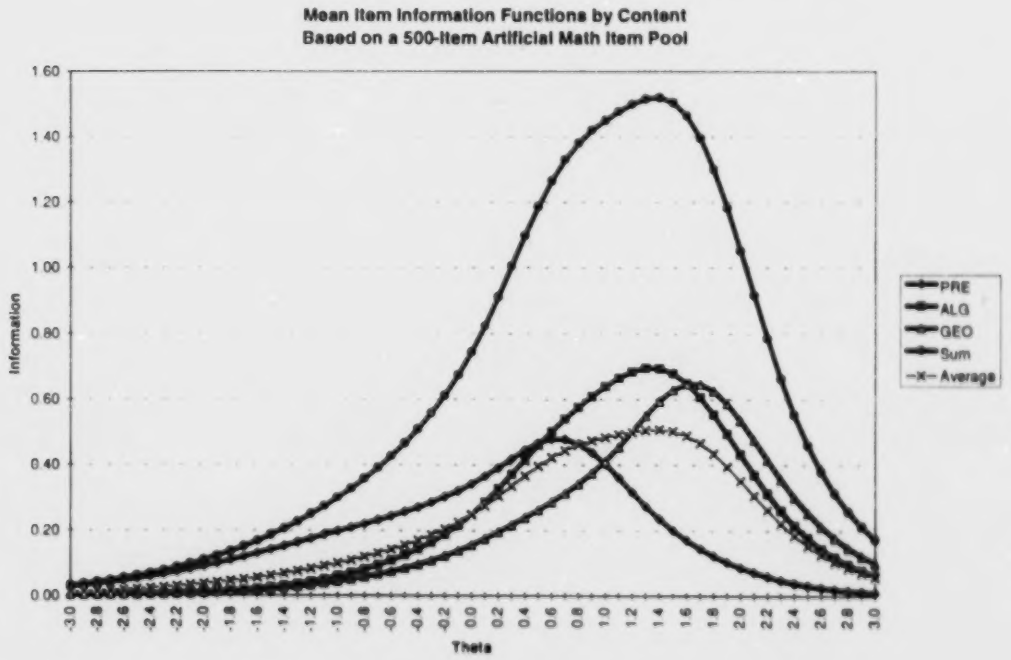

Figure 2. Average item information by content for a large, artificial Math CAT item pool

BEST COPY AVAILABLE

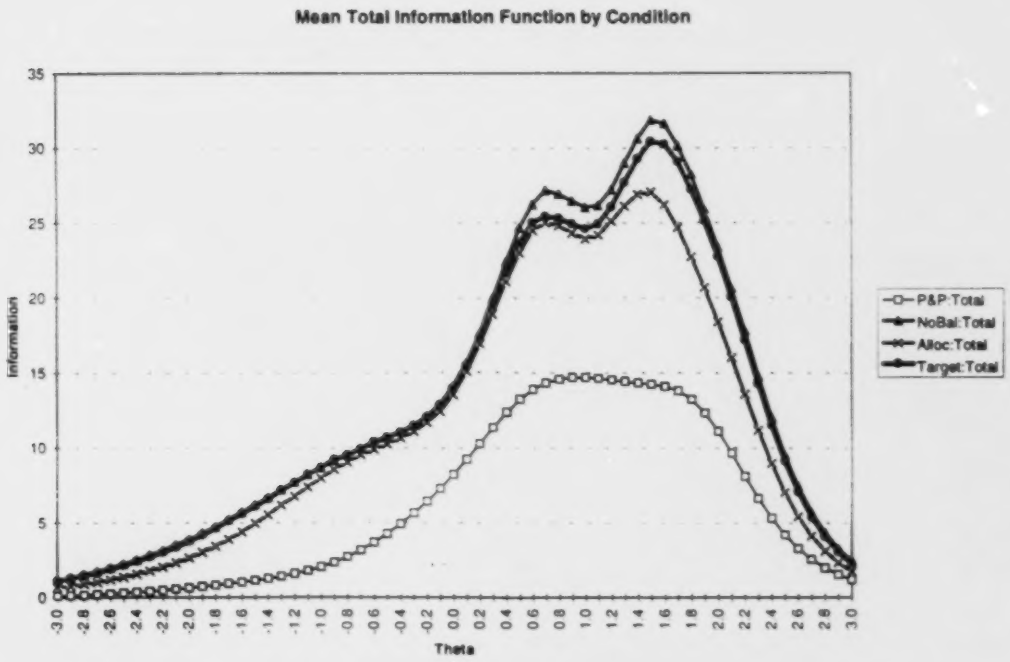

Figure 3. Mean total test information functions, for P&P target test and for each CAT condition

BEST COPY AVAILABLE

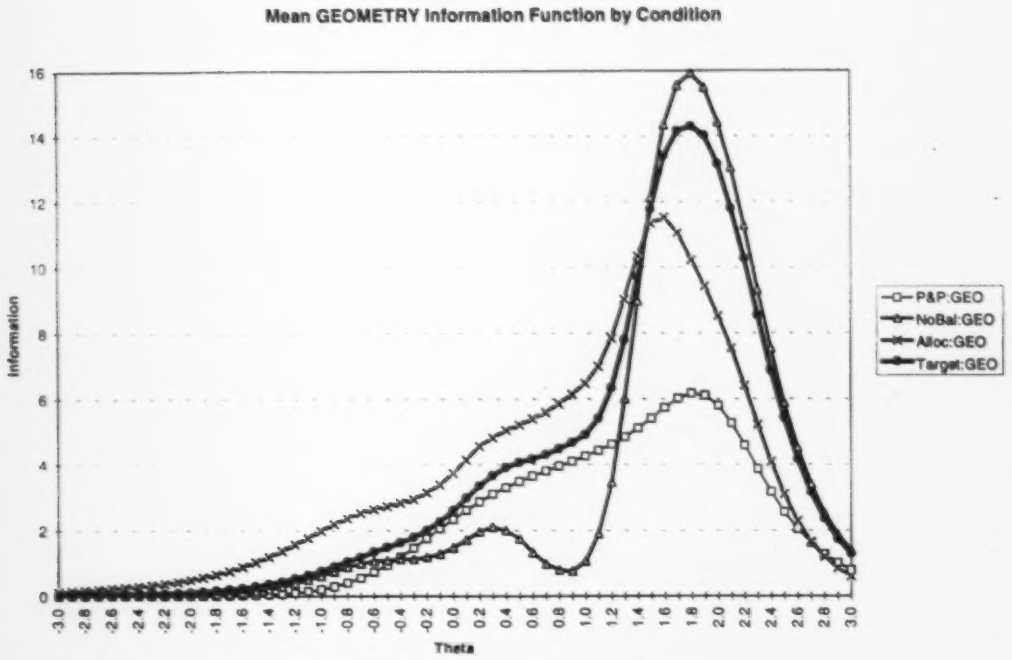

Figure 4. Mean Geometry test information functions, for P&P target test and for each CAT condition

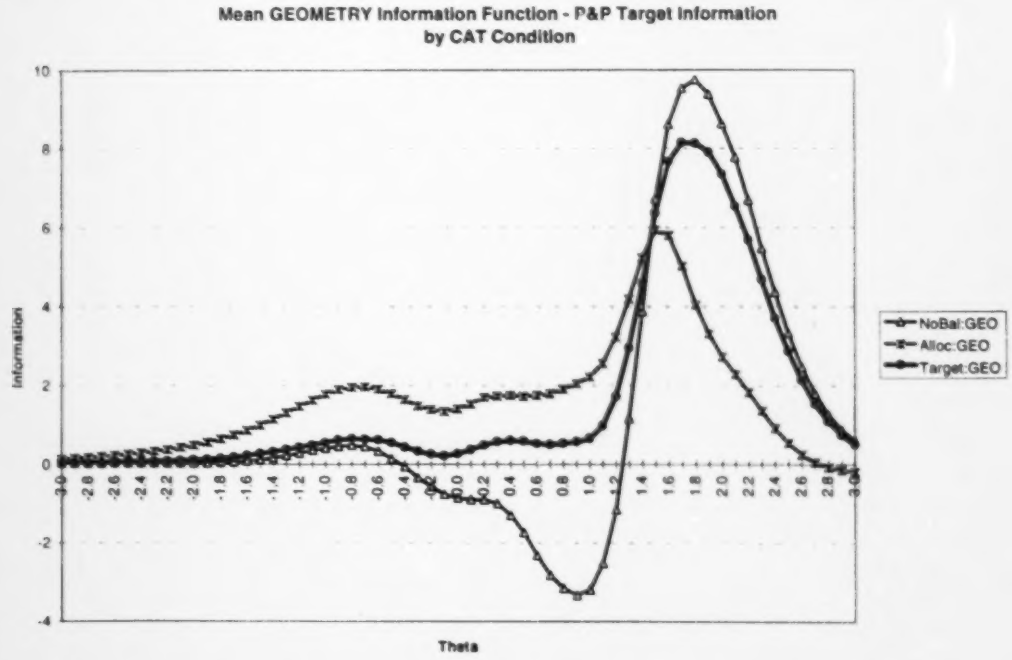

Figure 5. Mean Geometry information function minus P&P target information function for each CAT condition

BEST COPY AVAILABLE

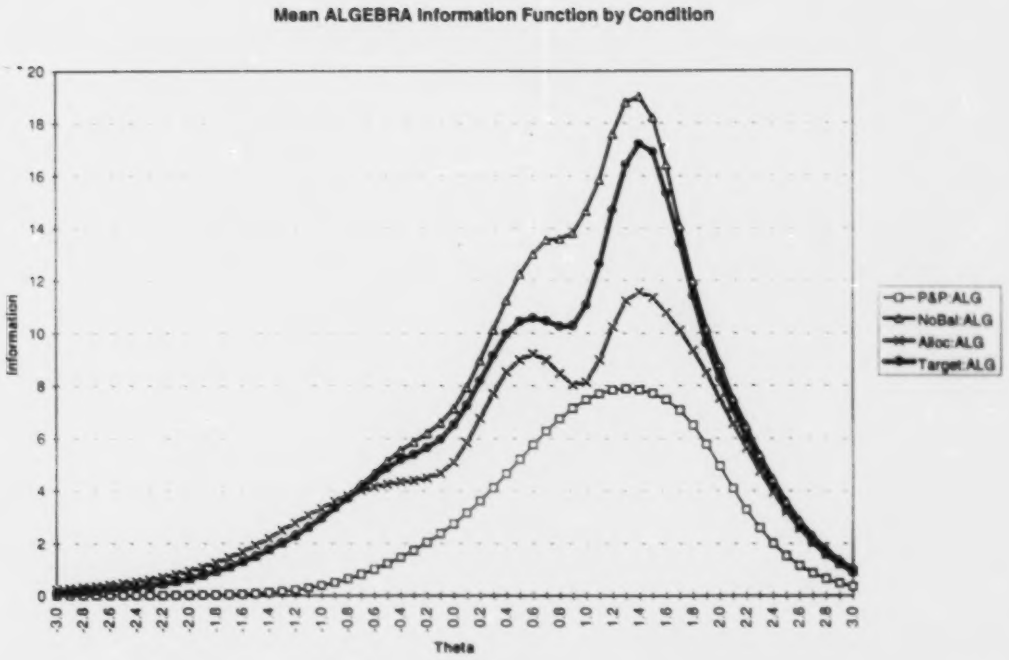

Figure 6. Mean Algebra test information functions, for P&P target test and for each CAT condition

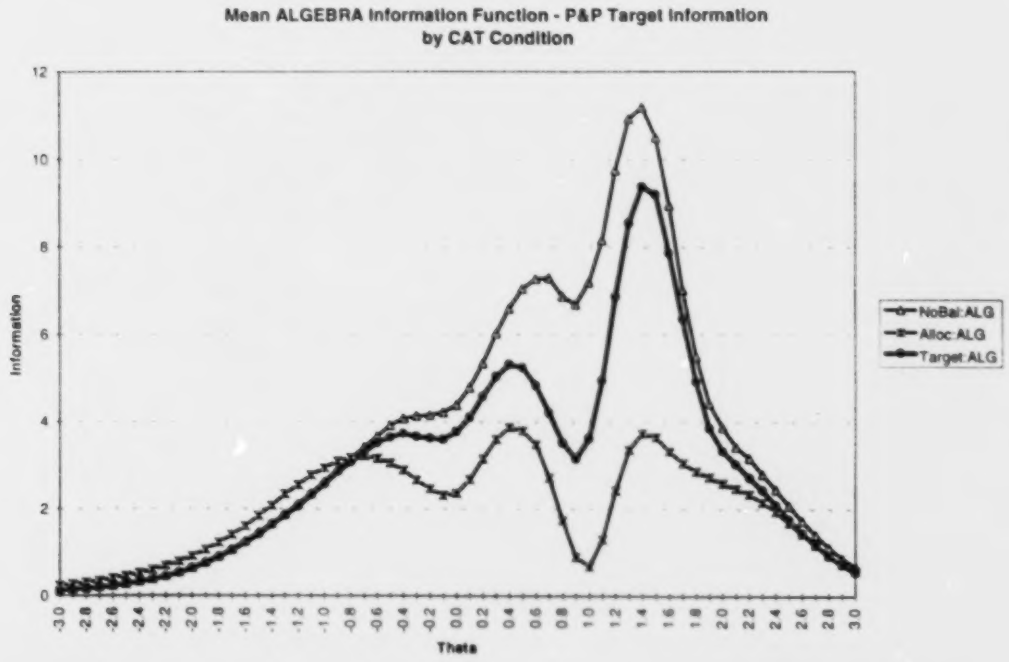

Figure 7. Mean Algebra information function minus P&P target information function for each CAT condition

BEST COPY AVAILABLE

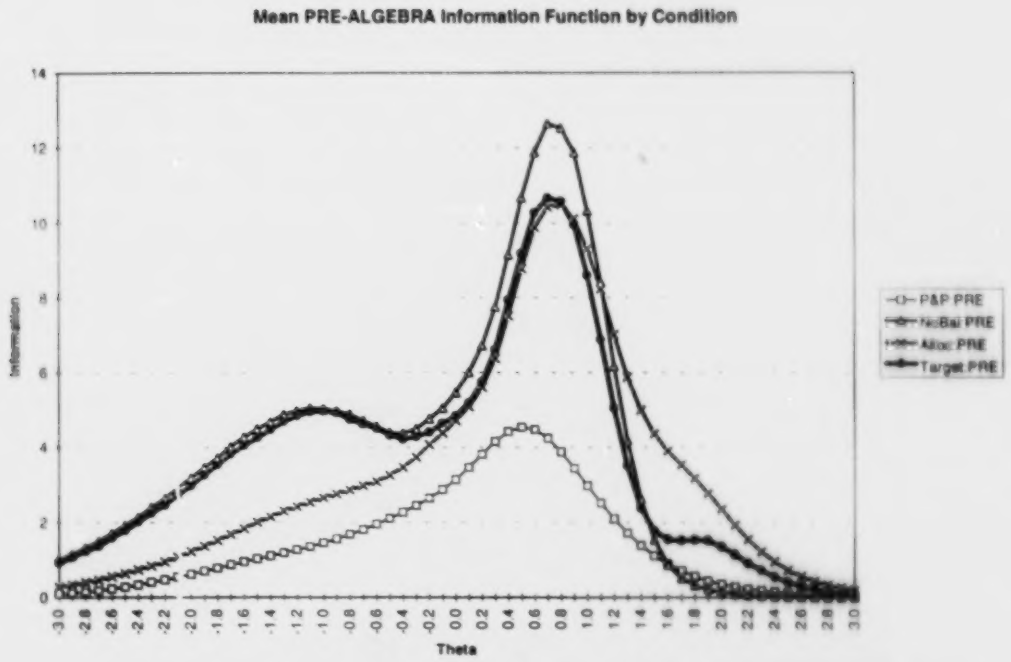

Figure 8. Mean Pre-algebra test information functions, for P&P target test and for each CAT condition

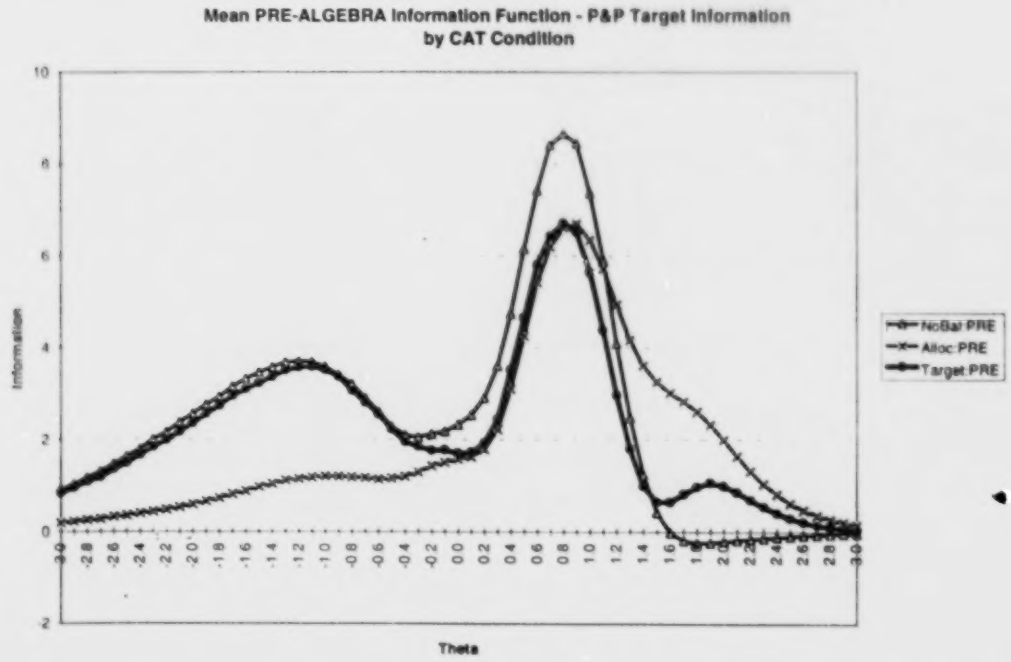

Figure 9. Mean Pre-algebra information function minus P&P target information function for each CAT condition

BEST COPY AVAILABLE

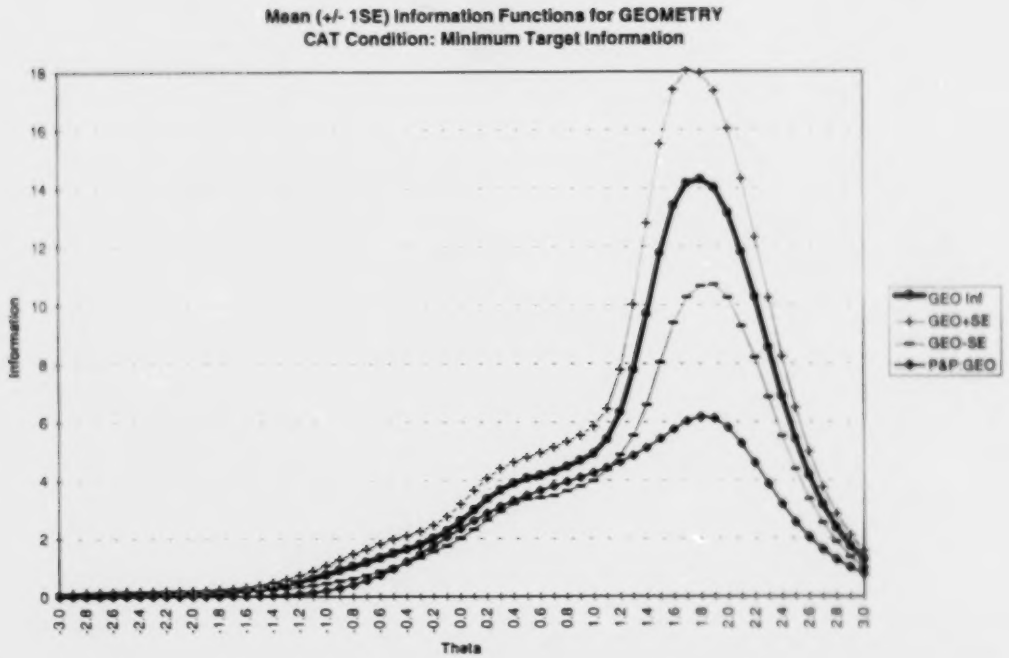

Figure 10. Mean Geometry test information functions plus and minus one SE for target CAT condition

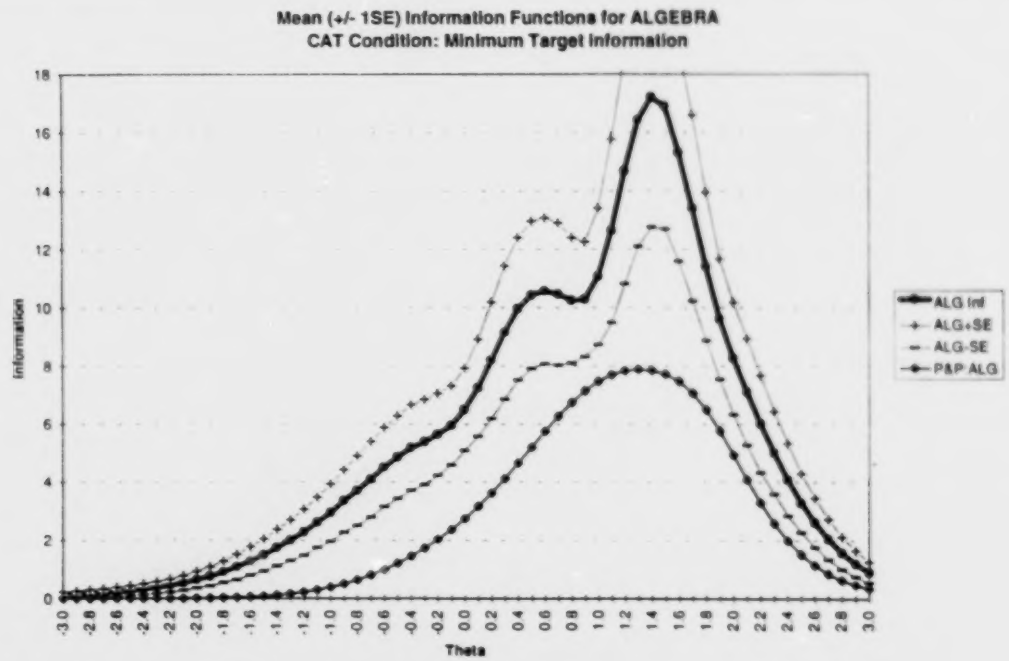

Figure 11. Mean Algebra test information functions plus and minus one SE for target CAT condition

BEST COPY AVAILABLE

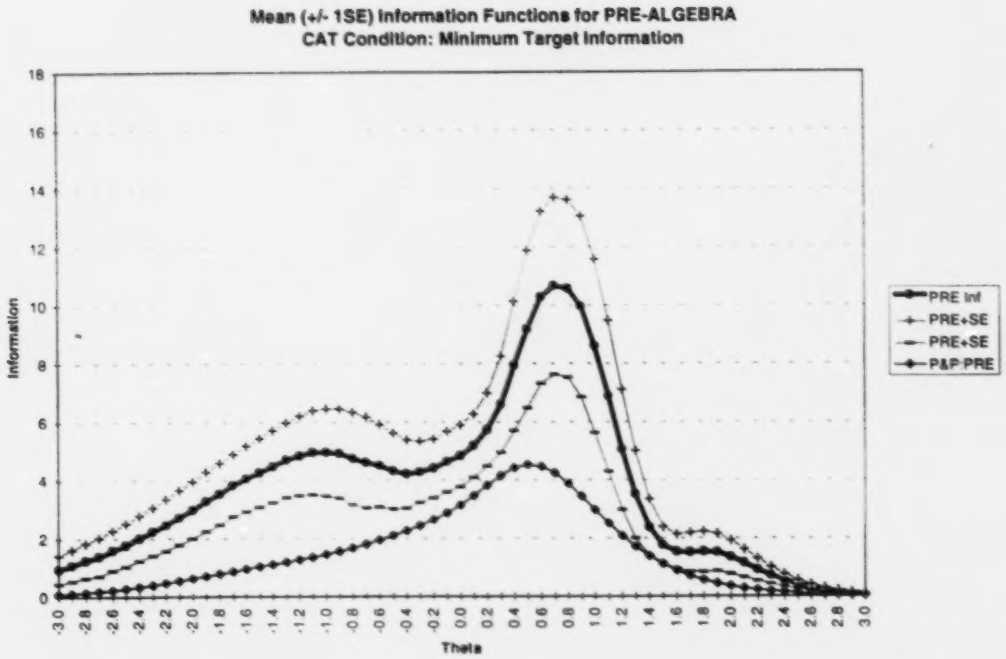

Figure 12. Mean Pre-algebra test information functions plus and minus one SE for target CAT condition

BEST COPY AVAILABLE

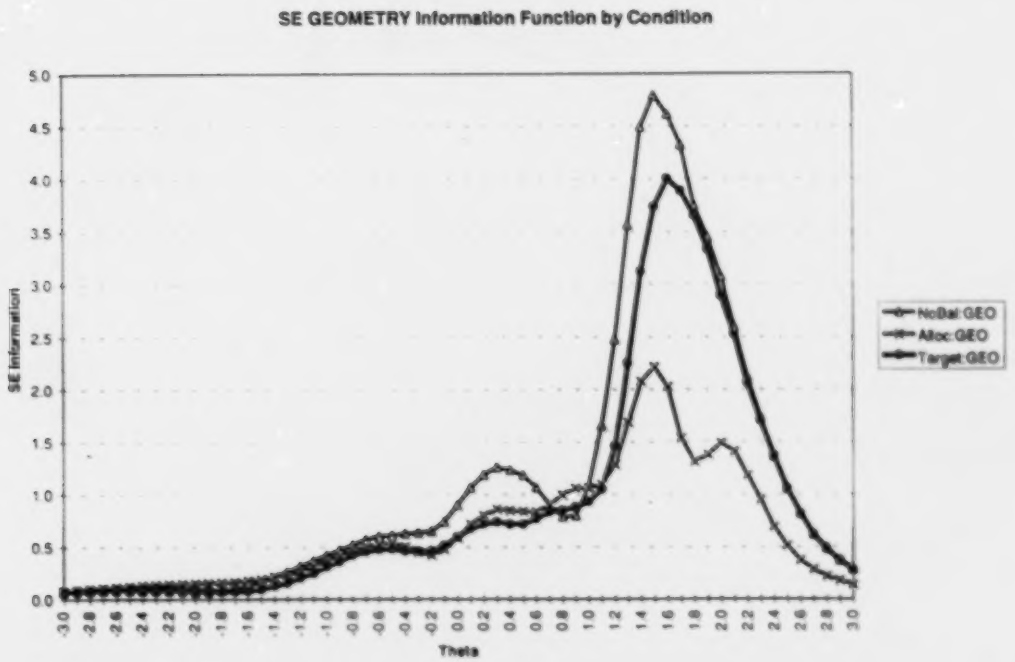

Figure 13. SE of Geometry test information functions for each CAT condition

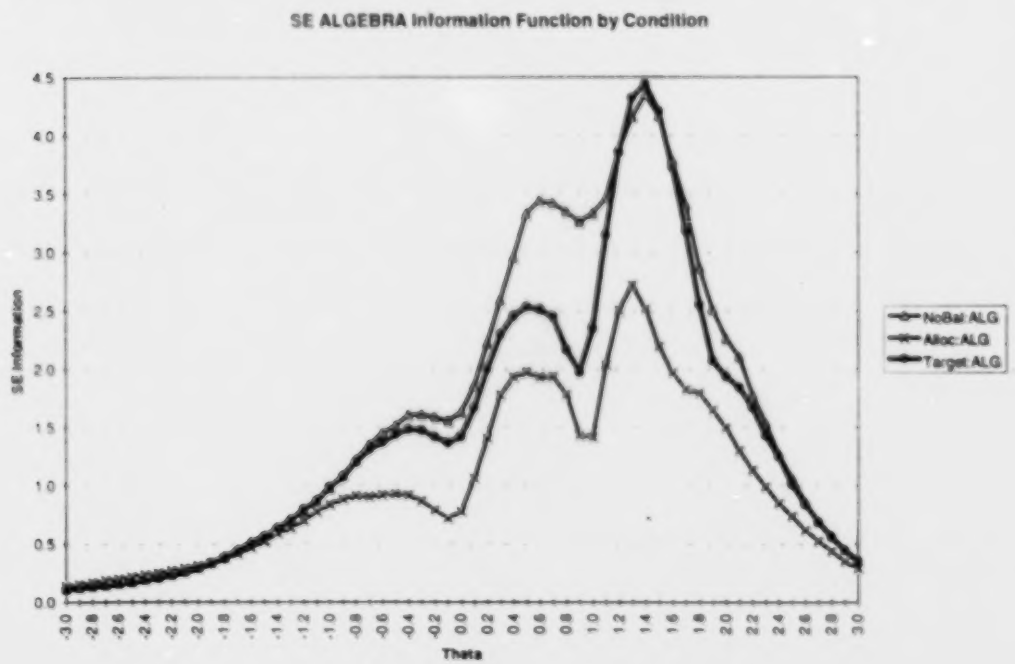

Figure 14. SE of Algebra test information functions for each CAT condition

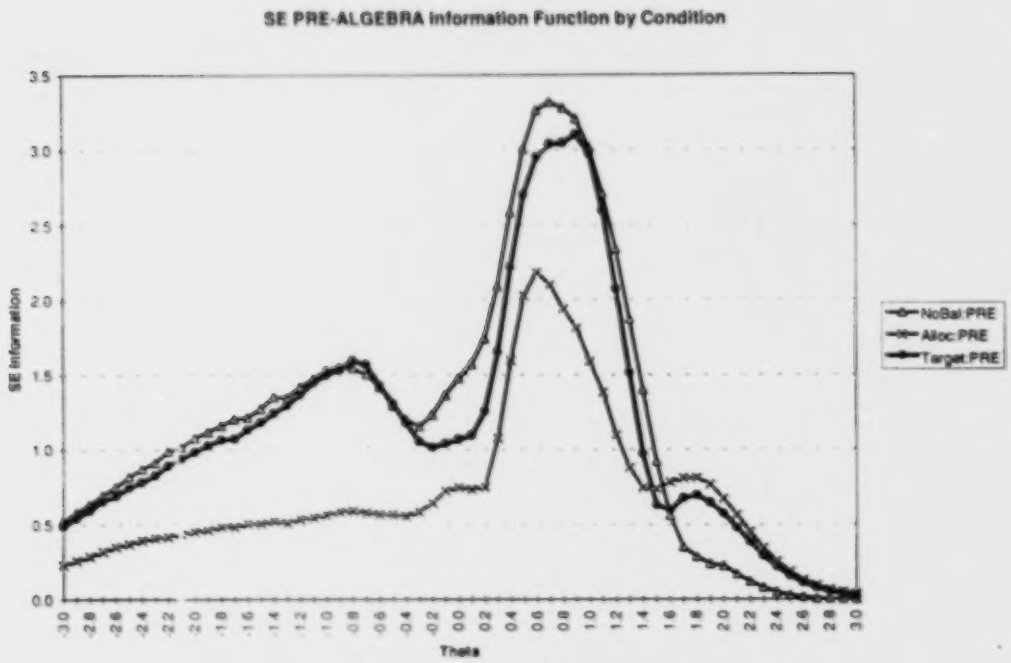

Figure 15. SE of Pre-algebra test information functions for each CAT condition

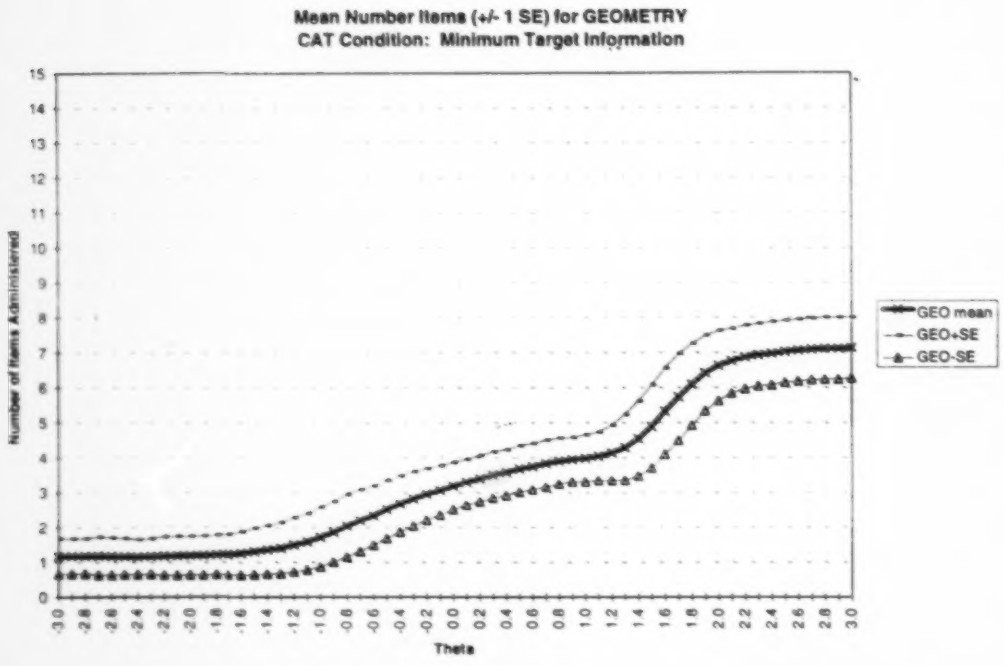

Figure 16. Mean number of administered Geometry items plus and minus one SE for target CAT condition

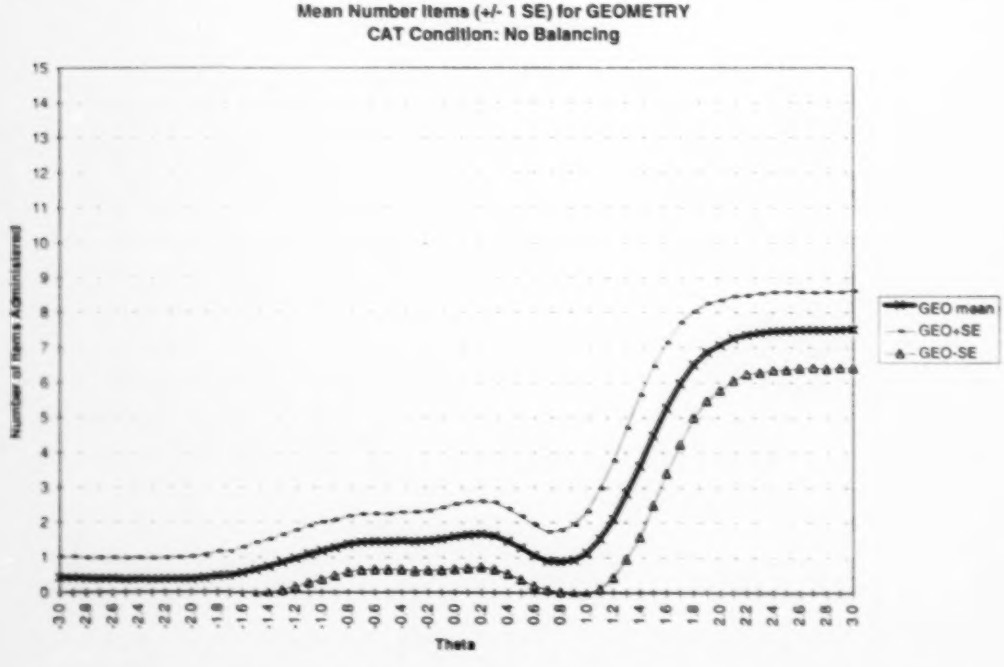

Figure 17. Mean number of administered Geometry items plus and minus one SE for no-balance CAT condition

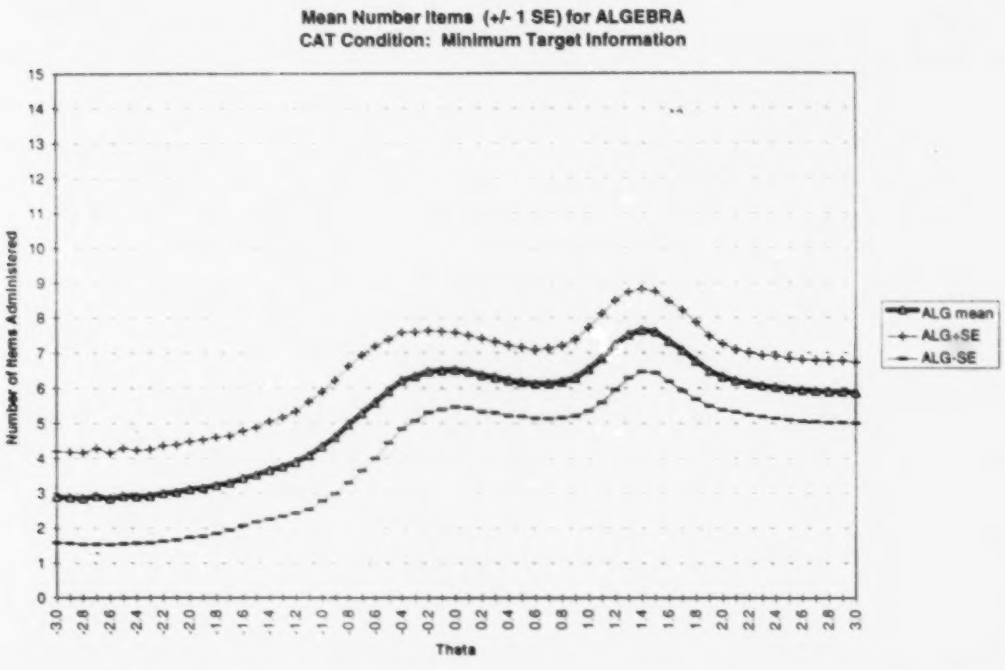

Figure 18. Mean number of administered Algebra items plus and minus one SE for target CAT condition

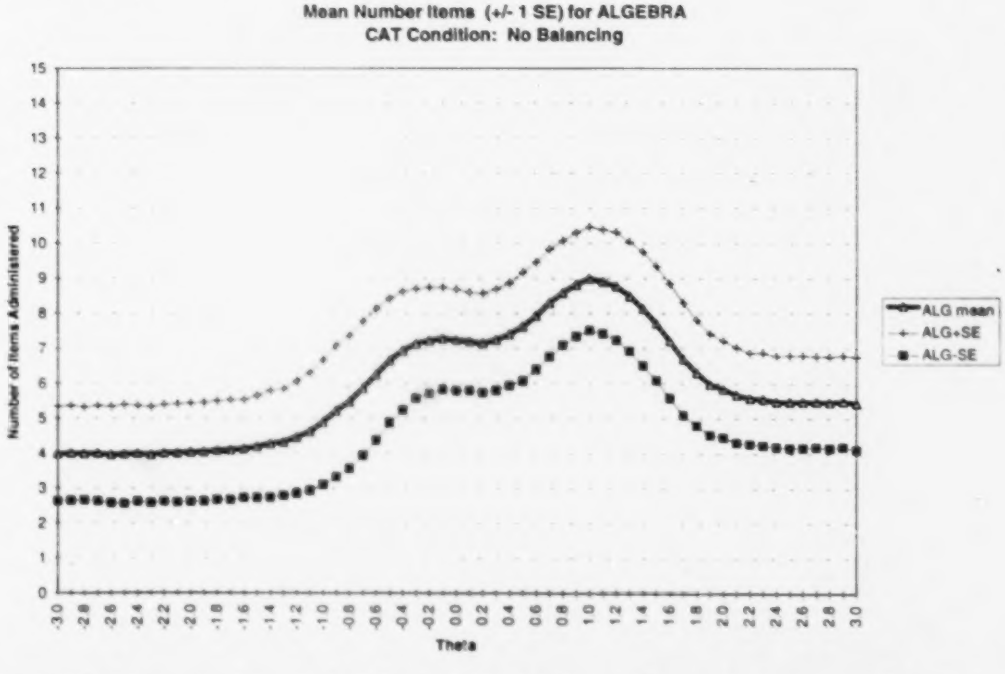

Figure 19. Mean number of administered Algebra items plus and minus one SE for no-balance CAT condition

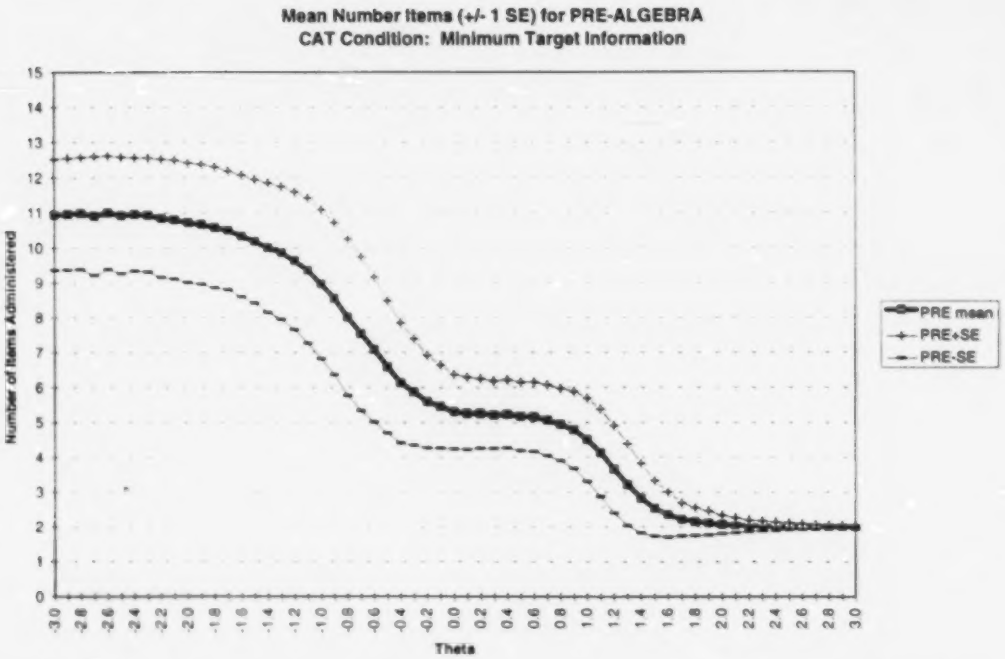

Figure 20. Mean number of administered Pre-algebra items plus and minus one SE for target CAT condition

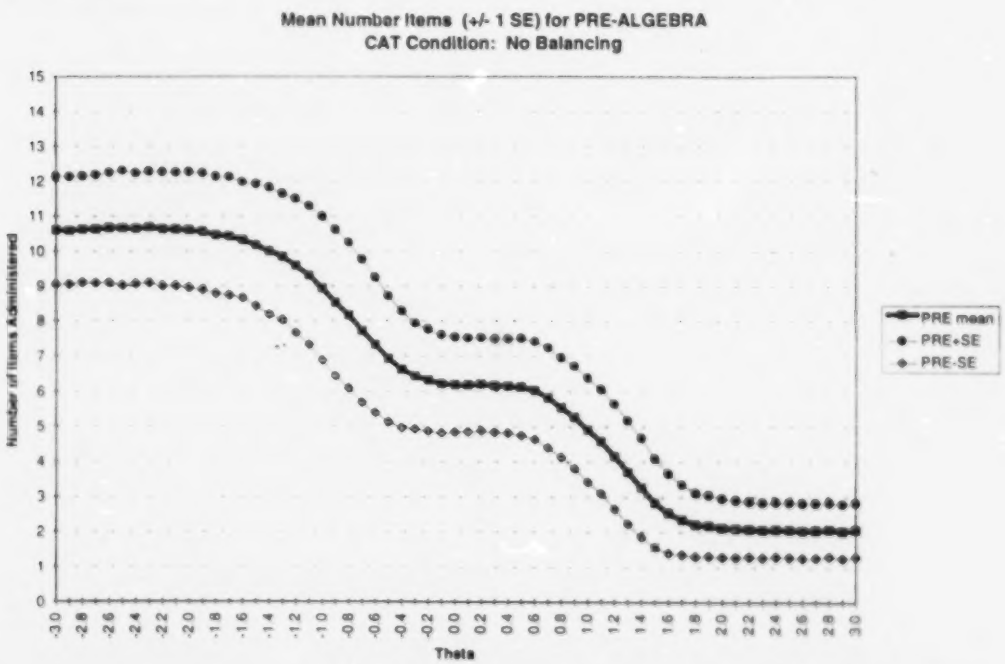

Figure 21. Mean number of administered Pre-algebra items plus and minus one SE for no-balance CAT condition

BEST COPY AVAILABLE